

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2001-344076
(P2001-344076A)

(43) 公開日 平成13年12月14日 (2001. 12. 14)

(51) Int.Cl. ⁷	識別記号	F I	テマコード* (参考)
G 0 6 F 3/06	3 0 5	G 0 6 F 3/06	3 0 5 C 5 B 0 1 8
	5 4 0		5 4 0 5 B 0 6 5
12/16	3 2 0	12/16	3 2 0 L 5 D 0 4 4
	3 4 0		3 4 0 P
G 1 1 B 20/10		G 1 1 B 20/10	H
審査請求 未請求 請求項の数 5 O L (全 21 頁) 最終頁に続く			

(21) 出願番号 特願2000-167484 (P2000-167484)

(22) 出願日 平成12年6月5日 (2000. 6. 5)

(71) 出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72) 発明者 森田 浩文

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(72) 発明者 石田 崇

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(74) 代理人 100100930

弁理士 長澤 俊一郎 (外1名)

最終頁に続く

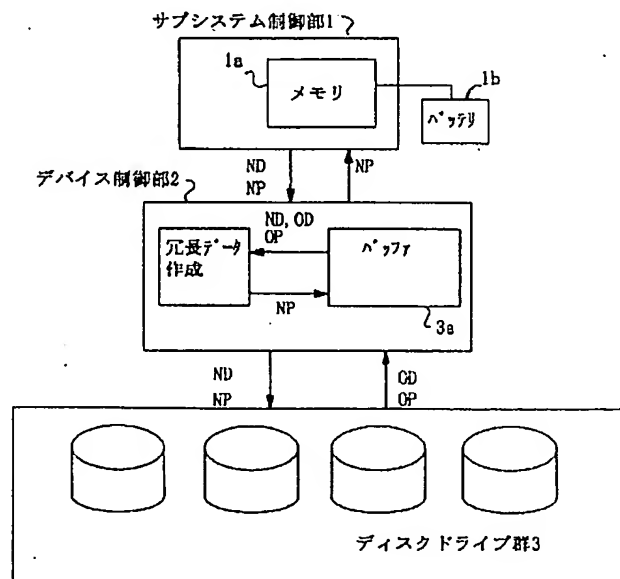
(54) 【発明の名称】 ディスクアレイ装置

(57) 【要約】

【課題】 データの信頼性を維持することができ、性能低下の問題が生ずることが少ないディスクアレイ装置を提供すること。

【解決手段】 R A I D 4, 5 のディスクアレイ装置において、ディスク縮退時、デバイス制御部2で作成した冗長データを、バッテリバックアップされたサブシステム制御部1のメモリ1aに転送し、書き込みが終了するまではメモリ上に冗長データを保持する。そして、電源瞬断の後の回復時、ディスクドライブ3からデータの読み込みを行わず、メモリ1a上の書き込みデータとパリティデータで書き込みを行う。また、ディスク正常時の電源瞬断の後の回復時、書き込み対象となるデータ及びパリティを格納するディスクドライブ以外のディスクドライブから冗長同一グループデータを読み出し、それらと書き込みデータから新冗長データを作成し対象ディスクへ書き込む。

本発明の概要を説明する図



【特許請求の範囲】

【請求項1】 少なくともメモリがバッテリバックアップされたサブシステム制御部と、

複数のディスクと、該ディスクを制御するデバイス制御部とを備え、

ライトデータを上記複数のディスクのそれぞれに分配して書き込むとともに、該分配された複数のデータから冗長データを生成し、該冗長データを上記複数のディスクのうち、分配したデータを格納したディスク以外のディスクに書き込むディスクアレイ装置であって、

上記ディスクヘデータを書き込む際、上記サブシステム制御部のメモリ上に書き込みデータを、書き込み処理が終了まで保持するとともに、上記複数のディスクの内の1乃至複数のディスクが故障したディスク縮退時には、デバイス制御部内のバッファに存在する冗長データをサブシステム制御部のメモリに転送して保持することを特徴とするディスクアレイ装置。

【請求項2】 ディスク縮退時に上記ディスクヘデータを書き込む際、書き込み対象となるディスクが故障している場合、ディスクへの書き込みデータと、正常動作しているディスク上の冗長同一グループのデータとから冗長データを生成し、該冗長データを上記サブシステム制御部のメモリに転送することを特徴とする請求項1のディスクアレイ装置。

【請求項3】 電源投入に際し、サブシステム制御部のメモリ中に保持された冗長データを使用して書き込み動作を行うことを特徴とする請求項1または請求項2のディスクアレイ装置。

【請求項4】 電源投入に際し、書き込み対象ディスク以外のディスクに格納された冗長同一グループのデータを用いて冗長データを生成することを特徴とする請求項1または請求項2のディスクアレイ装置。

【請求項5】 電源再投入時に、サブシステム制御部のメモリ内に未書き込みデータが存在するとき、請求項3または請求項4の書き込み動作を行うことを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、複数のディスク装置から構成され、ディスク故障時の修復機能を備えたディスクアレイ装置に関する。

【0002】

【従来の技術】近年、サブシステム内のコンポーネントを多重化し冗長度をもった装置が実用化されている。この装置は冗長構成と障害の極所化機能により、連続可用性を高めディスク一台の故障時におけるデータの自動修復機能を持つものであり、データの冗長化方法によりRAID0～RAID5までの6段階に分類される(RAID: Redundant Array of Inexpensive Disks)。図14(a)にRAID4のシステムの概略を示す。RAI

D4は同図に示すように、読み出し／書き込み単位に複数に分配されたデータをそれぞれ格納する複数のデータ用ディスクD0、D2、…と、パリティを生成する手段Pと、パリティを格納するディスク装置DPを備え、データの修復情報にはパリティ方式を採用している。

【0003】データは同図のA0、A2、…に示すように複数に分配され(一般には所定長)、分配されたデータA0、A2、…はそれぞれデータ用ディスクD0、D2、…に分散して格納され、パリティは一つの専用ディスクDPに格納される。なお、以下の説明では、上記のように分配されて異なったディスクに格納されたデータを冗長同一グループデータ、あるいは、単に冗長グループのデータと言い、これらのデータが格納されるディスク群を冗長グループのディスクと言う。また、パリティを冗長データともいう。ディスクに異常が発生した場合、ディスク上のデータは、残りの同一グループデータとパリティ(冗長データ)から再生される。RAID4は同時に複数の読み出しが可能だが、同時に複数の書き込みはできない。また、更新時には、必ず更新前のデータとパリティを読み出し、更新パリティを作成後書き込むといった余分なアクセスが必要である(これをライトペナルティという)。

【0004】図14(b)にRAID5のシステムの概要を示す。RAID5は、複数に分配されたデータとパリティを格納する複数のディスクD1、D2、…と、パリティを生成する手段Pとから構成され、RAID4と同様、データの修復情報にはパリティ方式を採用している。データは、同図のA0、A1、…、B0、B1、…に示すように複数に分配され、それぞれディスクD1、D2、…に分散して格納される。また、データA0、A1、…のパリティPA、データB0、B1、…のパリティPBも各ディスクディスクD1、D2、D3、…に分散して格納される。RAID5においても、RAID4と同様、ディスクに異常が発生した場合、ディスク上のデータは上記同一グループデータとパリティ(冗長データ)から再生される。RAID5は同時に複数ディスクの読み出し、書き込みが可能であり、更新時には、上記ライトペナルティが発生する。また、パリティ更新中のディスクへは、読み出し／書き込みのアクセスができない。

【0005】図15は上記RAID4又は5を適用可能なディスクアレイ装置における書き込みシーケンス例を示す図であり、同図はRAID4の場合を示している。同図において、101はサブシステム制御部、101aはサブシステム制御部内のメモリ、102はサブシステム内インタフェース部(以下、インタフェースをI/Fと略記する)、103はデバイス制御部、103aはバッファ、104はデバイスI/F部、105はディスク群であり、D0～D2はデータディスク、Pは冗長データを格納する冗長ディスクである。また、同図中のOD

(Old Data)は更新対象データ(以下では旧データともいう)、OP(Old Parity)は更新対象冗長データ(以下では、旧パリティともいう)、ND

(New Data)は書き込みデータ、NP(New Parity)は書き込み冗長データ(以下では新パリティともいう)、IPは中間冗長データである。

【0006】図15に示すようなディスクアレイ装置において、その書き込み動作は、以下のように行われる(図中の(a)~(g)は、以下の(a)~(g)に対応している)。

(a) 書き込みデータND1をサブシステム制御部101のメモリ101aから、デバイス制御部103のバッファ103aへ転送する。

(b) 書き込み対象となるディスク内のデータOD1をバッファ103aへ読み出す。

(c) 書き込み対象データの冗長グループの冗長データOPをバッファ103aへ読み出す。

(d) OD1とOPを排他的論理和演算することにより、中間冗長データIPを生成する。

(e) ND1とIPを排他的論理和演算することにより新冗長データNPを生成する。

(f) ND1をディスク105へ書き込む。

(g) NPをディスク105へ書き込む。

この際(a)~(c)及び(e)~(f)範囲内での順序は固定されなくとも書き込み動作は実現可能とされる。

【0007】

【発明が解決しようとする課題】上記のような書き込み動作を行うシステムにおいて停電等による瞬断が発生した場合のデータの信頼性の維持には以下のような方法が考えられている。

(1) 装置全体のバッテリーバックアップシステムによるサブシステムの動作の継続。

(2) 不揮発性メモリによる書き込みデータの維持
上記(1)においては装置に供給される電源が断たれた場合においてもサブシステムの動作は継続されるためそのデータは保証される。上記(2)においては書き込みデータがメモリ上に残存するため電源再投入後に再度ディスクへの書き込み動作を行うことで殆どの場合リカバリ可能である。しかしながら、上記(1)においてはサブシステム全体をバックアップするためには大容量なバッテリーを必要とし、実装上その占める割合は非常に大きなものとなる。また上記(2)においては電源断前に書き込みが行われている途中でかつ冗長データが書き込み途中であった場合で、そのRAID内が電源断前に縮退モード(少なくとも一台のディスクが故障時)にある、又は電源再投入後にRAID内の縮退モードに移行した場合には、その冗長グループの冗長性は失われ、故障ディスクのデータの復元又はその冗長グループのデータの書き込みは不可能(正しく行われぬ)となる。このような状態はWrite Holeとも表現されている。

【0008】このような問題に対応するため、メモリ101a上に冗長データを常時管理し、また書き込みの進行状況をメモリ101a上に記録しその進行状況によりその冗長データを使用してリカバリを行うようにし、上記状態を避けることも考えられてきているが、書き込み時に常に冗長データの転送が必要となるなどの性能低下が問題となる。本発明は上記事情に鑑みなされたものであって、本発明の目的は、縮退モードであったり、電源断の後の電源再投入時であっても、データの信頼性を維持することができ、性能低下の問題が生ずることが少ないディスクアレイ装置を提供することである。

【0009】

【課題を解決するための手段】図1は本発明の概要を説明する図である。同図に示すように、本発明は、少なくともメモリ1aがバッテリー1bによりバックアップされたサブシステム制御部1と、複数のディスクからなるディスクドライブ群3と、ディスクを制御するデバイス制御部2とを備え、ライトデータを上記複数のディスクのそれぞれに分配して書き込むとともに、該分配された複数のデータから冗長データを生成し、該冗長データを上記複数のディスクのうち、分配したデータを格納したディスク以外のディスクに書き込む前記RAID4、5のディスクアレイ装置において、以下のモードを設け、前記正常時、ディスク縮退時、電源瞬断後の回復時の書き込み時において、各状態に応じたモードで動作させる。

1) 正常時には、更新対象データODと更新対象冗長データOPと書き込みデータNDとから新冗長データNPを生成し、書き込みデータNDと新冗長データNPを対象ディスクへ書き込む。

2) ディスク縮退時には、データの書き込みの際し、デバイス制御部で作成した冗長データNPを、バッテリーバックアップされたサブシステム制御部のメモリに転送する。そして、書き込みが終了するまではメモリ上に冗長データNPを保持する。

3) 電源瞬断の後の再書き込み時、その対象となるデータ及びパリティを格納するディスクドライブ以外のディスクドライブから冗長同一グループ内のデータODを読み出し、それらとメモリ内の書き込みデータNDから新冗長データNPを作成し、書き込みデータNDと新冗長データNPを対象ディスクへ書き込む。

4) ディスク縮退時に、電源瞬断の後の再書き込みを行う場合、ディスクドライブからデータの読み込みを行わず、バッテリーバックアップされたメモリ上の書き込みデータNDとパリティデータNPで書き込みを行う。

以上のように、本発明においては、縮退モードのとき、冗長データNPを電源バックアップされたメモリ1aに転送するようにしたので、縮退モード時における電源瞬断後の再書き込み時にも、正しい書き込みを行うことができる。このため、常時、冗長データNPをメモリに転送する必要がなく、性能劣化を最低限に抑止でき、か

つ、高い信頼性を維持することができる。また、RAIDグループ内の他ディスクの冗長同一グループデータODと、メモリ内の書き込みデータNDから新冗長データNPを生成することにより、システム電源断後の電源再投入時、ディスク上のデータが正しくない、もしくは信用できない場合であっても、正しい書き込みを行うことができる。

【0010】

【発明の実施の形態】図2に本発明が適用可能なディスクアレイ装置の概略構成例を示す。図2は大規模なディスクアレイ装置に用いられる構成を示しており、11はサブシステム制御部であり、MPU、メモリ等を含む。12は上記サブシステム制御部11と他の制御部とを接続するための内部I/Fであり、複数のプロセッサを有するデバイス制御部13が接続される。デバイス制御部13は上記サブシステム制御部11の指示に従い、ディスクドライブ群15を制御する制御部であり、バッファ13a、MPU13b、デバイスI/F部14を制御するためのコントローラ、RAID4、5動作において必要な排他的論理和(Exclusive OR: XOR又はEORと記述されることもある)演算機能などを有する。14はデバイス制御部とディスクドライブを接続するためのデバイスI/F部であり、この構成例はシリアルI/Fのイメージであるが、パラレルI/Fであっても良い。15はデータが格納されるディスクドライブ群である。

【0011】図3は小規模なディスクアレイサブシステムに用いられる構成を示しており、21はサブシステムを制御するサブシステム制御部であり、サブシステム制御部21は図2に示したデバイス制御部の機能を兼ね、前記したメモリに加え、バッファ部21aとXOR演算機能を有する。14、15はそれぞれ図2に示したデバイスI/F部、ディスクドライブ群である。図4に図2に示す装置のハードウェア構成例を示す。同図において、サブシステム制御部11はチャンネルI/F部を介して上位装置に接続されており、サブシステム制御部11はメモリ11a、MPU11b、バスインタフェース部11cを備えており、上記MPU11bはメモリ11aに格納されているプログラムに従って動作する。また、メモリ11aには、プログラムの他に、転送データや制御データが格納される。上記サブシステム制御部11の少なくともメモリ11aはバッテリー11dによりバックアップされ、電源断時にも格納されたデータ等は保持される。なお、メモリ11aをバッテリーバックアップしても長時間の電源断には対応することができないので、例えば、図示しない特定のディスクドライブをバッテリーバックアップし、電源断時にメモリ11a内のユーザデータ及び制御情報(冗長データを含む)を上記ディスクに書き込み、電源再投入後にメモリ上に再度展開するように構成してもよい。

【0012】13はデバイス制御部であり、デバイス制御部13は、バッファ13a、MPU13b、上記MPU13bを動作させるプログラム等を格納したメモリ13c、バスインタフェース部13dを備えている。前記したXOR演算は、例えば上記MPU13aにより実行される。上記サブシステム制御部11とデバイス制御部13はバスBUSを介して接続されており、前記図2に示したサブシステム内I/F部12は、上記バスBUS、バスインタフェース部11c、バスインタフェース部13d等から構成され部分に対応する。デバイス制御部13は前記したようにデバイスI/F部14を介してディスクドライブ群15に接続される。

【0013】次に、図2に示すディスクアレイ装置を前述のRAID4、5を適用した場合について、本実施例の動作を説明する。なお、以下の説明ではRAID4について説明するが、RAID5の場合についても同様に適用することができる。また、以下の実施例では、図2に示すシステムについて説明するが、本発明は図3のシステムにも同様に適用することができる。図2、図4において、本実施例のサブシステム制御部11はディスクドライブ群内で構成されるRAIDの状態を正常/縮退/データ喪失(同一グループのデータの内、2つのディスクのデータが読めないもしくは信用できない場合)に区別して個々に管理し、その状態を参照してディスクドライブ群への書き込み命令を実行する。

【0014】また、このサブシステム制御部11はデバイス制御部12に書き込みを指示する際、以下のモードを指定することができ、デバイス制御部12はそれを実行する機能を持つ。

(1) モード1

更新対象データODと、更新対象冗長データOPを読み出し、排他的論理和をとり中間冗長データIPを生成し、さらに、書き込みデータNDとの排他的論理和をとり、書き込み冗長データNPを生成し、NPとNDの対象ディスクへの書き込みを行う。この「モード1」は正常時に選択され動作は前記図15で説明した通りである。

(2) モード2

デバイス制御部13がNPを作成した際、それをサブシステム制御部11のメモリ11a上へ転送し、転送後に対象ディスクへの書き込みを行う。このモード2は、後述するようにディスク縮退状態の書き込み時に選択される。

【0015】(3) モード3

OD、OPをディスクから読み出すことなく、冗長同一グループの他のディスク内の上記ODと同一のアドレスからデータを読み出しそれらとメモリ上のNDからNPを生成し、対象ディスクへの書き込みを行う。このモード3は、後述するように、ディスクが縮退状態でないとき、書き込み中に予期しない電源断が生じ、電源再投入後にサブシステム制御部11のメモリ11aに未書き込

みデータが存在している場合の再書き込み時に選択される。

(4) モード4

OD、OPをディスクから読み出すことなく、メモリ上にあるNPとNDの対象ディスクへの書き込みを行う。このモード4は、後述するようにディスク縮退時、書き込み中に予期しない電源断が生じ、電源再投入後にサブシステム制御部11のメモリ11aに未書き込みデータが存在している場合の再書き込み時に選択される。

【0016】上記モードの切り替えは次のように行われる。ディスクの一台が故障し、デバイス制御部13からサブシステム制御部11が異常終了通知を受けると、サブシステム制御部13は、特定ディスクへのアクセスが不能で動作継続不可能である場合、縮退状態と認識し、書き込み時のモードを正常モードから縮退モードに移行する。また、サブシステム制御部11の書き込み命令に対するデバイス制御部13からの正常応答の有無を判断し正常応答が無いままサブシステム制御部11が装置電源断を認識した場合、その書き込み処理を仕掛かり中と判断し、電源再投入時に、未書き込みデータがメモリ11a上に存在している場合、上記モード3（ディスク正常時）、モード4（ディスク縮退時）で書き込みを行う。

【0017】図5に、上記正常時のモード（モード1）から縮退時のモード（モード2）への移行のフローチャートを示す。図5において、上位装置からRAIDへのアクセスがあり、正常終了であれば、正常モードのまま次のRAIDへのアクセスを待つ。正常終了でない場合には、特定ディスクへのアクセスが不能であるのかを調べる。特定ディスクへのアクセスが不能である場合には、縮退モードに移行する。また、特定ディスクへのアクセスが不能でない場合には、リトライを行い、正常終了すれば正常モードのまま、次のRAIDへのアクセスを待つ。正常終了しない場合には、所定回数に達するまでリトライを行い、所定回数に達すると、エラー処理を行う。

【0018】次に、本実施例のディスクアレイシステムの動作について説明する。

(1) 正常時

対象となるRAIDが正常な状態ではサブシステム制御部は、前記「モード1」での書き込み動作をデバイス動作に対し指示を行う。この場合の動作は、図15で説明したのと同じである。図6に、正常時における通常の書き込み動作のフローチャートを示す。同図において、まず、書き込みデータND1をメモリ11aに保持したまま、バッファ13aに転送する（ステップS1）。ついで、書き込み対象となっている旧データ（更新対象データ）OD1をディスク15からバッファ13aに読み込む（ステップS2）。次に、書き込み対象の冗長データ（更新対象冗長データ）OPをディスク15からバッ

ファ13aへ読み込む（ステップS3）。

【0019】そして、OD1とOPの排他的論理和（XOR）を演算し、中間パリティ（中間冗長データ）IPを作成し（ステップS4）、さらに、ND1とIPの排他的論理和（XOR）を演算し、新パリティ（書き込み冗長データ）NPを作成する（ステップS5）。以上のようにして新パリティNPが作成されたら、書き込みデータND1、新パリティNPのディスク15への書き込みを開始する（ステップS6）。そして、書き込みが終了するまで待ち（ステップS7）、書き込みが正常終了したらメモリ11a上のND1を削除して（ステップS9）処理を終了する。また、書き込みが正常終了しなかった場合には、ステップS8からステップS10に行き、リトライ回数が所定回数になったかを調べ、所定回数になっていなければ、ステップS6に戻り、書き込みをリトライする。そして、所定回数リトライしても正常終了しない場合にはエラー処理を行う（ステップS11）。

【0020】（2）ディスクドライブ群のあるディスクドライブに異常が発生した場合（ディスク縮退時）

ディスクドライブ群のあるディスクドライブに何らかの異常が発生した場合は図7に示すように動作する。サブシステム制御部11からの読出し／書き込み又はその他のディスクアクセス系のコマンドに対し、ディスクドライブ群15のあるディスクドライブ（例えば図7におけるディスクD1）に何らかの異常が発生した場合、デバイス制御部13はサブシステム制御部11へ、異常終了通知を行う（図7の(a)）。通知を受けたサブシステム制御部11は、その内容を判断してディスクの故障による動作継続不可能である場合、内部の構成変更情報の変更（RAID状態を縮退状態に変更する）を行い（図7の(b)）、そのディスクドライブを切り離す命令を出し（図7の(c)）、縮退モードに移行する。

【0021】この縮退モードにあるRAIDに対する書き込み動作を図8のフローチャートにより説明する。書き込みデータND2をサブシステム制御部11のメモリ11aに保持したまま、書き込みデータND2を、デバイス制御部13のバッファ13aに転送する（ステップS1）。ついで、ND2が故障したディスクに書き込むデータであるかを調べ（ステップS2）、故障ディスクに書き込むデータでない場合には、書き込み対象となっている旧データ（更新対象データ）OD2をディスク15からバッファ13aに読み込み（ステップS3）、更新対象冗長データ（OP）をディスク15からバッファ13aに読み込む（ステップS4）。ついで、旧データOD2と更新対象冗長データOPの排他的論理和（XOR）を作成しIPとし（ステップS5）、書き込みデータND2とIPの排他的論理和（XOR）を作成しNPとする（ステップS6）。

【0022】また、ND2が故障したディスクに書き込

むデータである場合には、更新対象データOD2の冗長同一グループデータOD0、OD1をディスク15から読み込み（ステップS7）、OD0とOD1と書き込みデータND2との排他的論理和を求めて新パリティNPとする（ステップS8）。上記のようにして新パリティNPが作成されたら、NPをサブシステム制御部11のメモリ11aに転送する（ステップS9）。ついで、正常終了するまで待ち、書き込みデータND2、新パリティNPのディスク15への書き込みを開始する（ステップS10、S11）。そして、書き込みが終了するまで待ち（ステップS12）、書き込みが正常終了したらメモリ11a上のND2を削除して処理を終了する（ステップS14）。また、書き込みが正常終了しなかった場合には、ステップS13からステップS15に行き、リトライ回数が所定回数になったかを調べ、所定回数になっていなければ、ステップS11に戻り、書き込みをリトライする。そして、所定回数リトライしても正常終了しない場合にはエラー処理を行う（ステップS16）。以上のように、ディスク縮退時には、作成した新冗長データNPをサブシステム制御部11のメモリ11aに転送し保持する（前記モード2の動作）。

【0023】図9に、上記した書き込みデータND2が故障したディスクに書き込むデータでない場合の動作シーケンス例を示す。サブシステム制御部11は書き込み動作をデバイス制御部13に指示する前に該当するRAIDの状態を確認する。この場合、その書き込み対象となるRAIDの状態は縮退とされているのでサブシステム制御部11はデバイス制御部13に対して次のように書き込みを指示する。

- (a) 書き込みデータND2をメモリ11aに保持したまま、メモリ11aからバッファ13aへ転送する。
- (b) 書き込み対象となるディスク内の旧データOD2をバッファ13aへ読み出す。
- (c) 書き込み対象部分が属する冗長グループの冗長データOPをバッファ13aへ読み出す。
- (d) OD2とOPをXOR演算することにより、中間冗長データIPを生成する。
- (e) ND2とIPをXOR演算することにより新冗長データNPを生成する。
- (f) NPをサブシステム制御部11のメモリ11aに書き込む。
- (g) ND2、NPをディスク15へ書き込み、メモリ11a上のND2を削除する。

【0024】図10に上記した書き込みデータND2が故障したディスクに書き込むデータの場合の動作シーケンス例を示す。

- (a) 書き込みデータND2をメモリ11aに保持したまま、メモリ11aからバッファ部13aへ転送する。
- (b) (c) OD2、OPをディスクから読み出すことなく、冗長グループの他のディスク内の旧データOD0、

OD1をバッファ部13aへ読み出す。

(d) OD0、OD1、ND2をXOR演算することにより、新冗長データNPを生成する（中間冗長データIPの作成については同図では省略されている。以下同様にIPの作成についての説明は省略する）。

(e) NPをサブシステム制御部11のメモリ11aに書き込む。

(f) ND2、NPを書き込み対象ディスク15へ書き込み、メモリ11a上のND2を削除する。

10 【0025】(3) 電源断後の回復時の動作

次に、書き込み動作中に瞬断等による予期せぬ電源断の場合を考える。前記「モード1（正常時のモード）」で、書き込み動作中に予期せぬ電源断が生じた際、サブシステム制御部11のメモリ11aの電源のみがバックアップされているとした場合、ディスクドライブ群15への対象ディスクに対する書き込みは中断される。すなわち、サブシステム制御部11の書き込み命令に対するデバイス制御部からの正常応答の有無を判断し正常応答が無いままサブシステム制御部11が装置電源断を認識した場合、その書き込み処理を仕掛かり中と判断し、対象ディスクに対する書き込みは中断される。その後、電源が再投入された後にメモリ11a上に未書き込みのデータが存在している場合には書き込みは再試行される。ここで、ディスク縮退状態でない場合（正常時）であっても、再度「モード1（正常時のモード）」で書き込みを行った場合、電源前の書き込みは中断しているため旧データOD（更新対象データ）の読み込み時のデータは正しくない可能性があり使用することはできない。そこで電源投入時に未書き込みのデータが存在している場合、サブシステム制御部11は前記「モード3」で書き込みを行うようにデバイス制御部13に指示を出す。

【0026】また、ディスクが縮退状態にあるとき、その書き込みは前記「モード2」で行われる。「モード2」で書き込み動作中に、予期せぬ電源断が生じた場合、サブシステム制御部11のみ電源がバックアップされているとした場合、前記したようにディスクドライブ15のそのディスクに対する書き込みは中断される。その後、電源が再投入された後にメモリ11a上に未書き込みのデータが存在している場合には書き込みは再試行されるが、このとき再度「モード2」で書き込みを行った場合、電源前の書き込みは中断しているため、更新対象データ（旧データ）ODの読み込み時のデータは正しくない可能性があり使用することはできない。さらに、前記「モード3」で書き込みを再試行した場合、そのRAIDはすでに縮退状態にあるためParityの生成は不可能となる。しかし、瞬断前の書き込みは前記「モード2」で行われているためメモリ内に未書き込みの新冗長データNPが保持されている。そこで電源投入時にメモリ11a内に未書き込みのデータおよび未書き込みのNPが存在している場合には、サブシステム制御部1

1は前記「モード4」で書き込みを行うようデバイス制御部13に指示を出す。

【0027】上記電源瞬断後の回復時の書き込み動作について、図11のフローチャートにより説明する。電源瞬断後の回復時、メモリ11a上に書き込みデータND2が登録されているかを調べ（ステップS1）、登録されていない場合には処理を終了する。また、登録されている場合には、ND2の属する冗長グループが縮退モードになっているかを調べる（ステップS2）。縮退モードになっていない場合には、書き込み対象となっている更新対象データOD2と冗長同一グループのデータOD0、OD1をディスク15からバッファ13aに読み込む（ステップS3）。また、メモリ11aからバッファ13aに書き込みデータND2を転送する（ステップS4）。ついで、上記OD0とOD1と書き込みデータND2の排他的論理和（XOR）を作成し、新冗長データNPとする（ステップS5）。一方、縮退モードの場合には、ステップS6において、メモリ11a上に新冗長データNPが保持されているかを調べ、保持されている場合には、書き込みデータND2、NPをバッファ13aに転送する。また、保持されていない場合にはステップS3に行く。ついで、書き込みデータND2、新パリティNPのディスク15への書き込みを開始する（ステップS8）。

【0028】そして、書き込みが終了するまで待ち（ステップS9）、書き込みが正常終了したらメモリ11a上のND2を削除して処理を終了する（ステップS10、S11）。また、書き込みが正常終了しなかった場合には、ステップS10からステップS12に行き、リトライ回数が所定回数になったかを調べ、所定回数になっていなければ、ステップS8に戻り、書き込みをリトライする。そして、所定回数リトライしても正常終了しない場合にはエラー処理を行う（ステップS13）。

【0029】図12に縮退状態モードでない場合における電源瞬断後の回復時の書き込み動作シーケンス例を示す。

- (a) 書き込みデータND2をメモリ11aからバッファ部13aへ転送する。
- (b) (c) 冗長グループ内のOD2と同一のアドレスから冗長同一グループのデータOD0、OD1をバッファ部13aへ読み出す。
- (d) 上記冗長同一グループデータOD0、OD1、ND2をXOR演算し、新冗長データNPを作成する。
- (e) ND2、NPをディスクへ書き込む。

【0030】図13に縮退モードにおける電源瞬断後の回復時の書き込み動作シーケンス例を示す。

- (a) 書き込みデータNDをメモリ11aからバッファ部13aへ転送する。
- (b) NPをメモリ11aからバッファ部13aへ転送する。

(c) NDをディスクへ書き込む。

(d) NPをディスクへ書き込む。

【0031】

【発明の効果】以上説明したように、本発明をRAID4、5が適用可能なディスクアレイ装置に適用することにより以下の効果を得ることができる。

(1) 縮退モードのときに冗長データを電源バックアップされたメモリに転送するようにしたので、従来例のように常時、冗長データをメモリに転送する場合と比べ、性能劣化を最低限に抑止でき、かつ、高い信頼性を維持することができる。

(2) 縮退モードにおいて、システム電源断後の電源再投入時の書き込み動作時にメモリ上の冗長データを使用することでディスク上への冗長データの書き込み途中で有る無いに問わず正しい書き込みを行うことができる。

(3) RAIDグループ内の他ディスクの同一グループデータと、メモリ内の書き込みデータから新冗長データを生成することにより、縮退時、あるいは、ディスク上のデータが書き込み途中で有る無いに問わず正しい冗長データを作成することができる。

特に、ディスク正常時のシステム電源断後の電源再投入時に、RAIDグループ内の他ディスクの同一グループデータと、メモリ内の書き込みデータから新冗長データを生成することにより、ディスク上のデータが正しくない、もしくは信用できない場合であっても、正しい書き込みを行うことができる。

【図面の簡単な説明】

【図1】本発明の概要を説明する図である。

【図2】本発明が適用可能なディスクアレイ装置の概略構成例を示す図（1）である。

【図3】本発明が適用可能なディスクアレイ装置の概略構成例を示す図（2）である。

【図4】本発明の実施例のハードウェア構成例を示す図である。

【図5】縮退モードの設定処理のフローチャートである。

【図6】正常時における書き込み動作のフローチャートである。

【図7】ディスクドライブ群のあるディスクドライブに何らかの異常が発生した場合の動作を説明する図である。

【図8】縮退モードのときの書き込み動作を示すフローチャートである。

【図9】縮退モードの書き込み動作シーケンス例（1）を示す図である。

【図10】縮退モードの書き込み動作シーケンス例（2）を示す図である。

【図11】電源瞬断後の回復時の書き込み動作を示すフローチャートである。

【図12】電源瞬断後の回復時の書き込み動作シーケンス例（縮退モードでないとき）を示す図である。

【図13】電源瞬断後の回復時の書き込み動作シーケンス例（縮退モード時）を示す図である。

【図14】RAID4, 5を説明する図である。

【図15】ディスクアレイ装置における書き込み動作シーケンス例を示す図である。

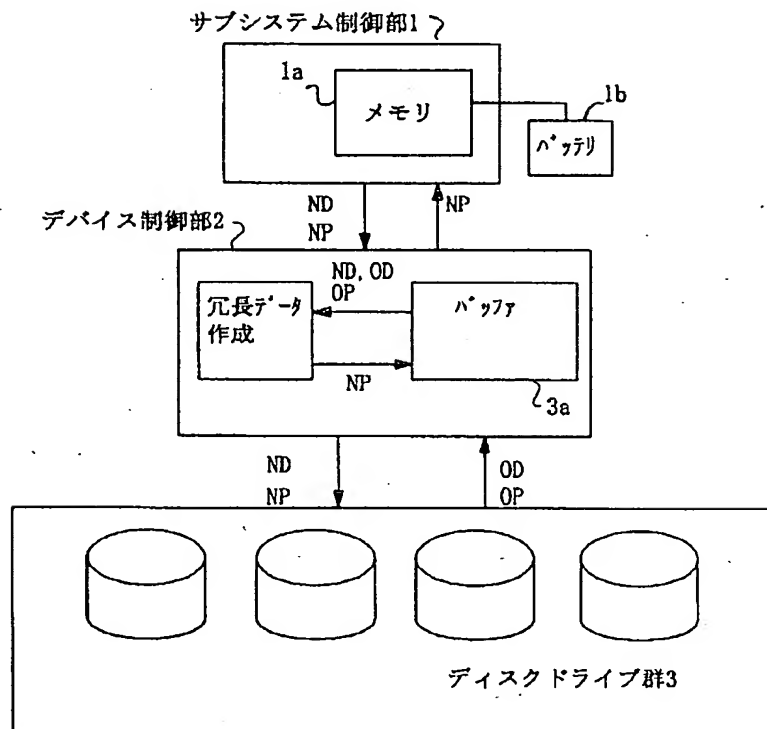
【符号の説明】

1 サブシステム制御部
1 a メモリ
1 b バッテリ

2 デバイス制御部
3 ディスクドライブ群
11 サブシステム制御部
11 a メモリ
11 b MPU
12 サブシステム内インタフェース部
13 デバイス制御部
13 a バッファ
13 b MPU
10 14 デバイスインタフェース部
15 ディスクドライブ群

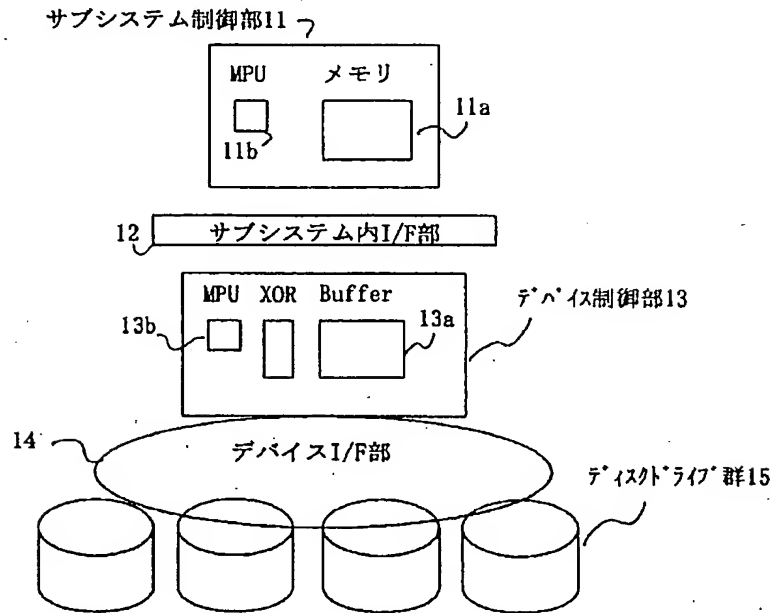
【図1】

本発明の概要を説明する図



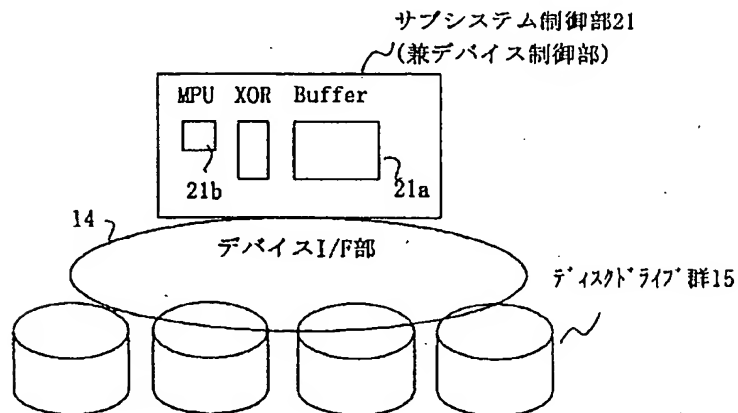
【図2】

本発明が適用可能なディスクアレイ装置の概略構成例（１）
を示す図



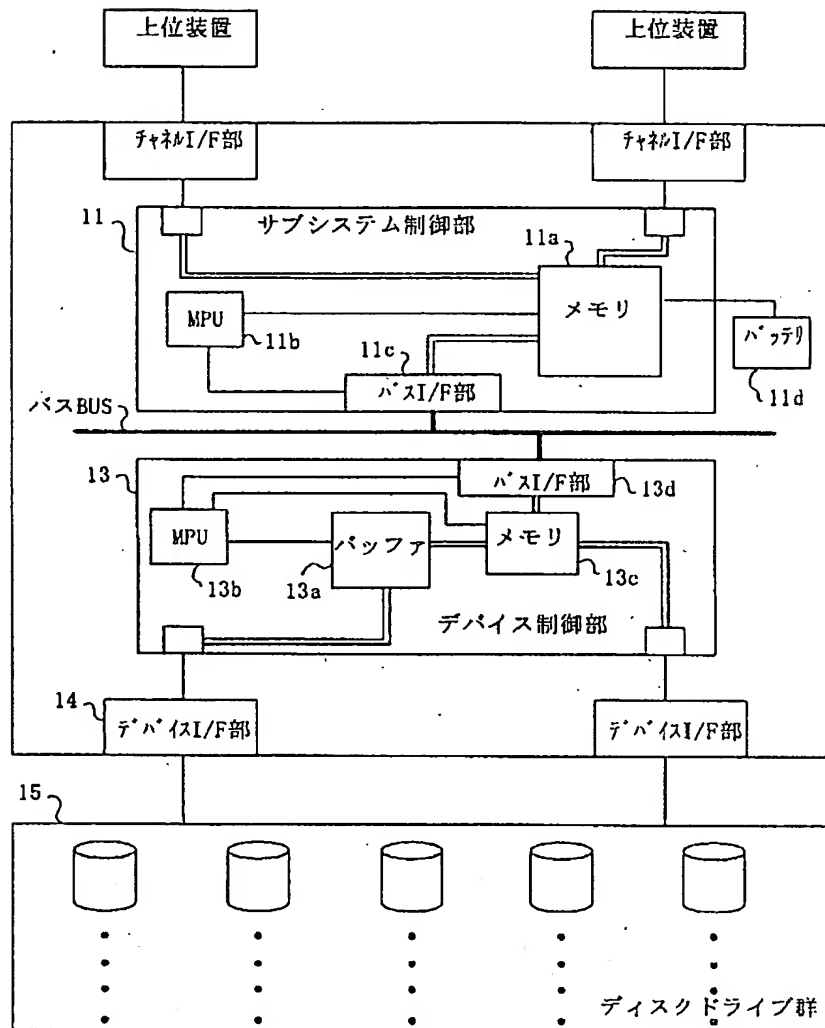
【図3】

本発明が適用可能なディスクアレイ装置の概略構成例（２）
を示す図



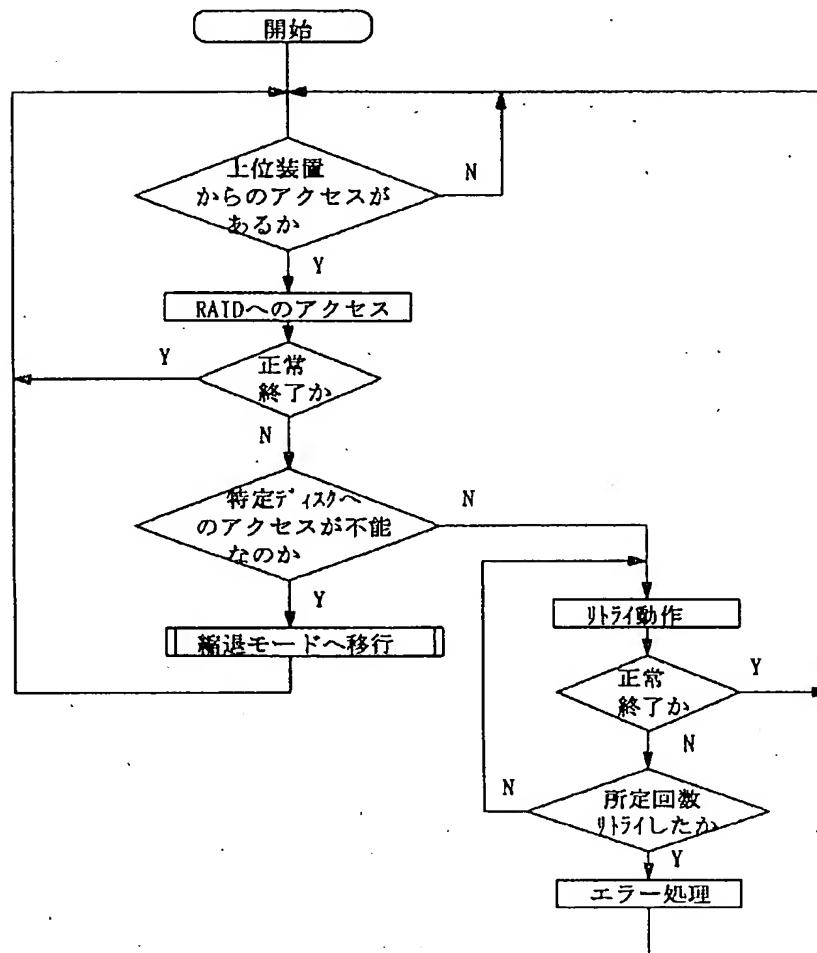
【図4】

本発明の実施例のハードウェア構成例を示す図



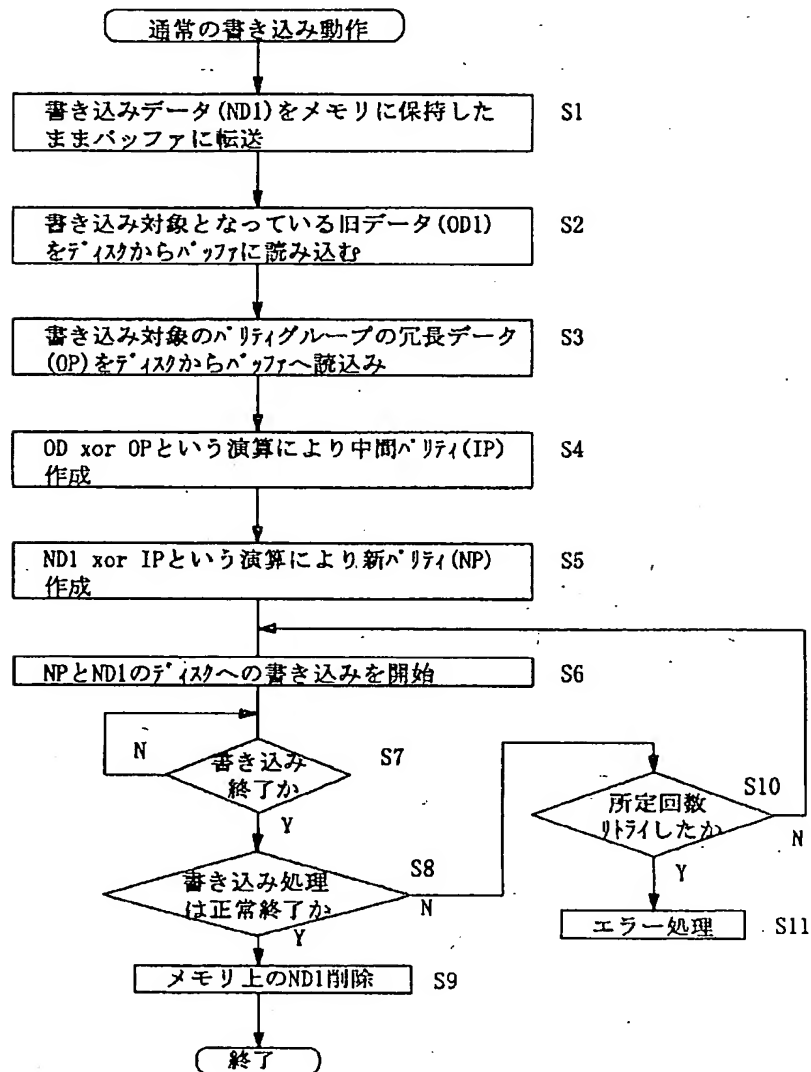
【図5】

縮退モードの設定処理のフローチャート



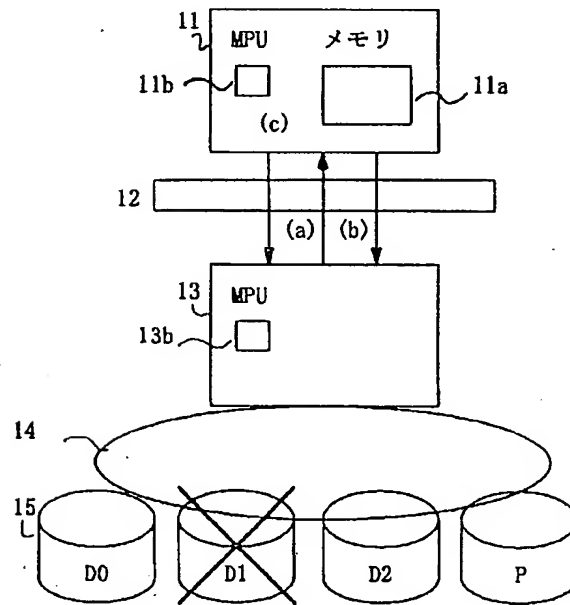
【図6】

正常時における書き込み処理のフローチャート



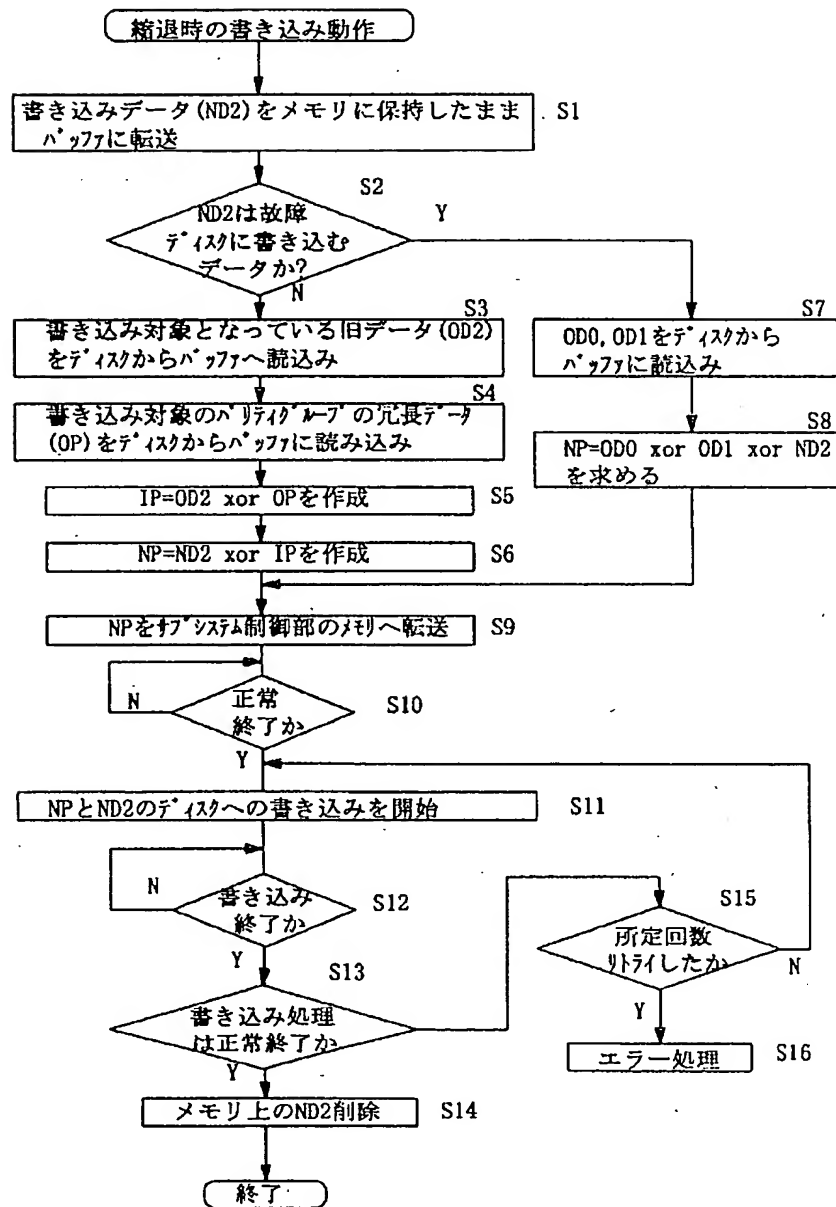
【図7】

ディスクドライブ群のあるディスクに何らかの異常が発生した場合の動作を説明する図



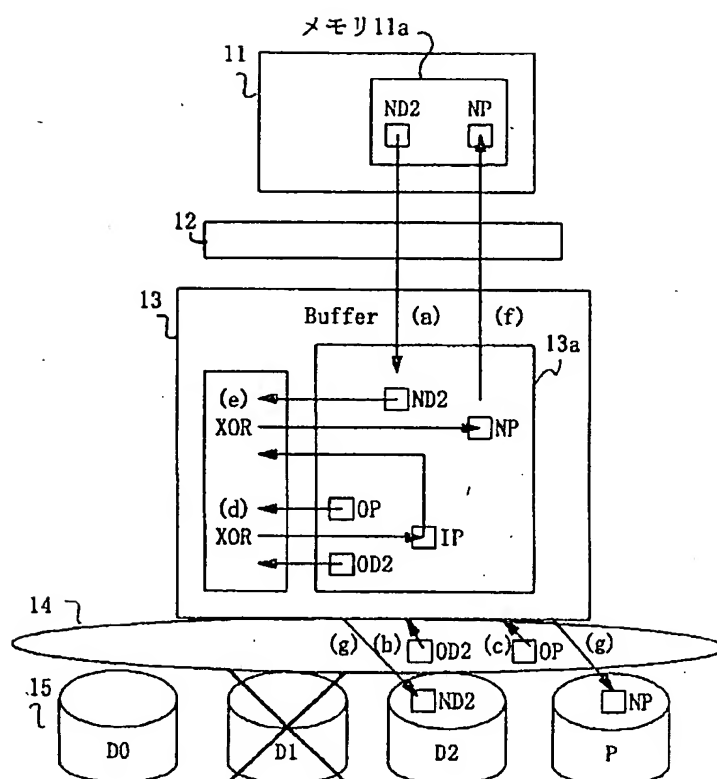
【図8】

縮退モードのときの書き込み動作を示すフローチャート



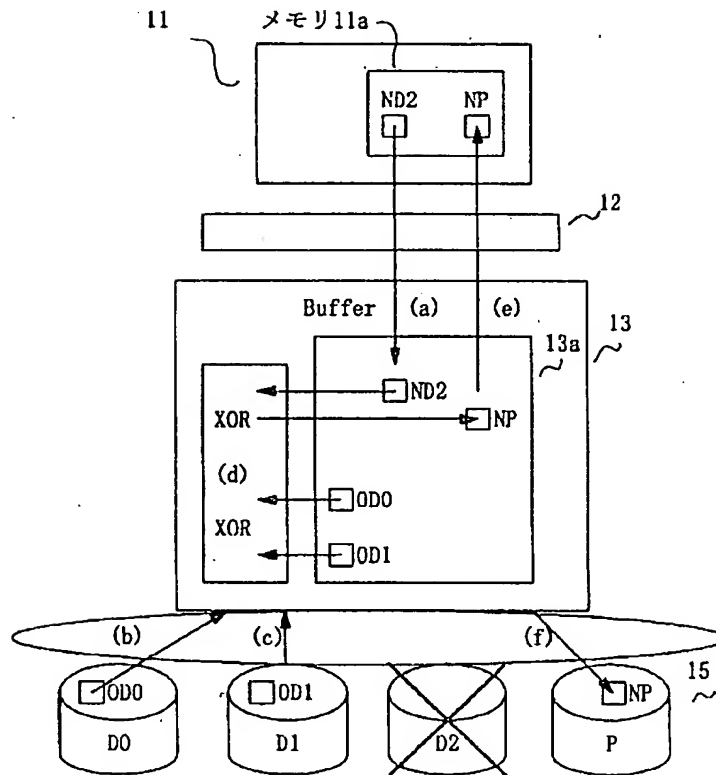
【図9】

縮退モードの書き込み動作シーケンス例（1）を
示す図



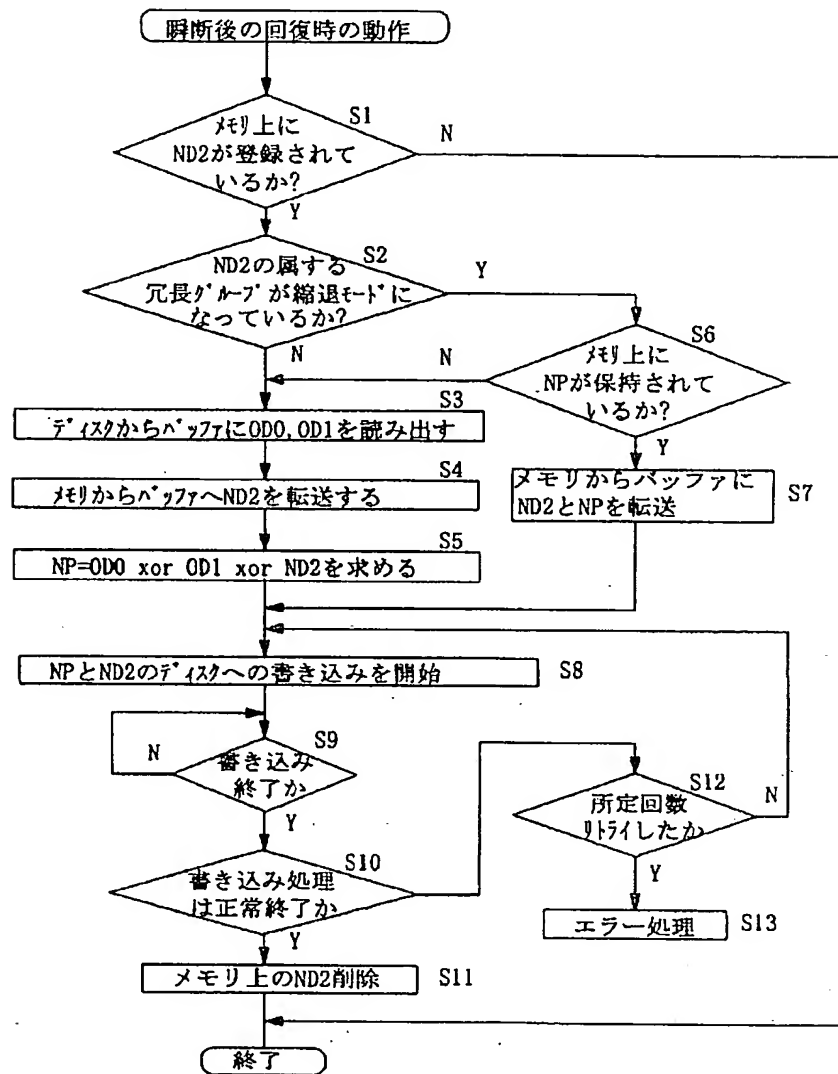
【図10】

縮退モードの書き込み動作シーケンス例（２）を示す図



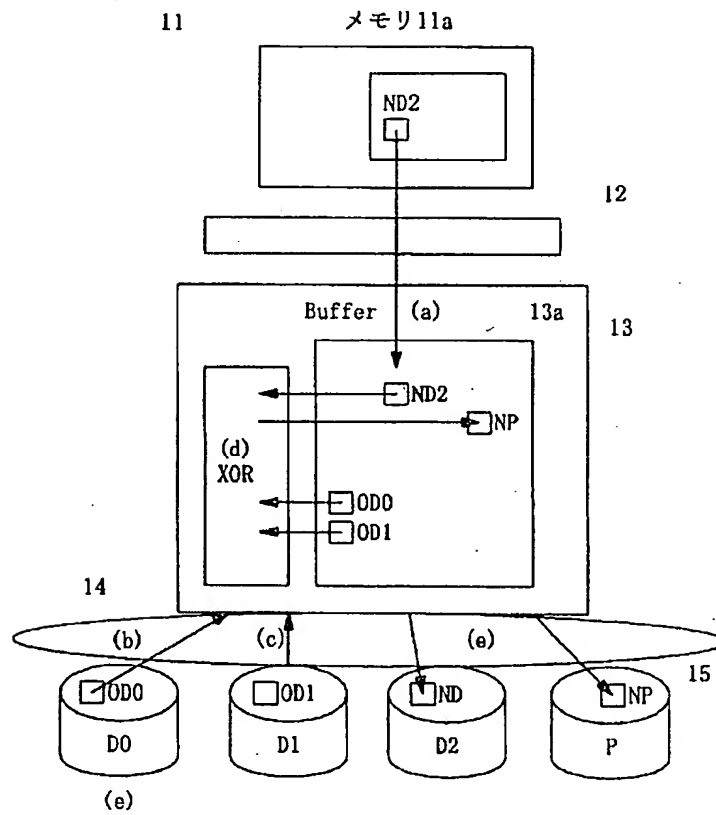
【図11】

電源瞬断後の回復時の書き込み動作を示すフローチャート



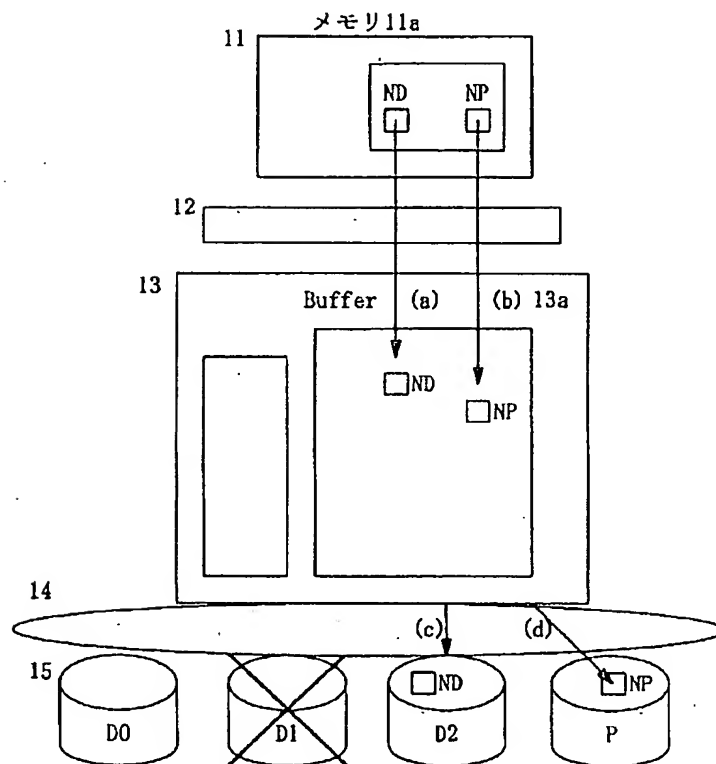
【図12】

電源瞬断後の回復時の書き込み動作シーケンス例 (1)
を示す図



【図13】

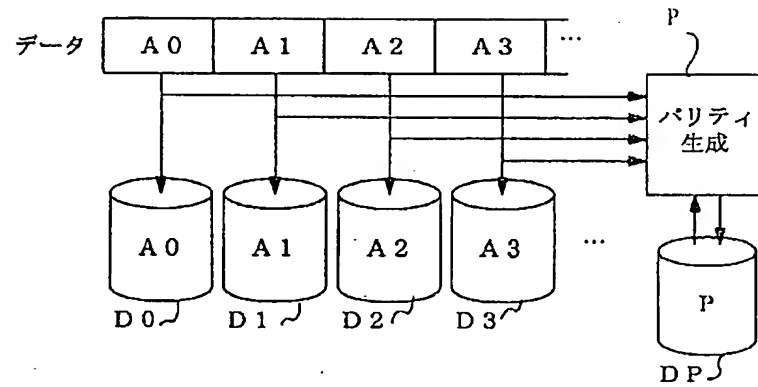
電源瞬断後の回復時の書き込み動作シーケンス例（2）
を示す図



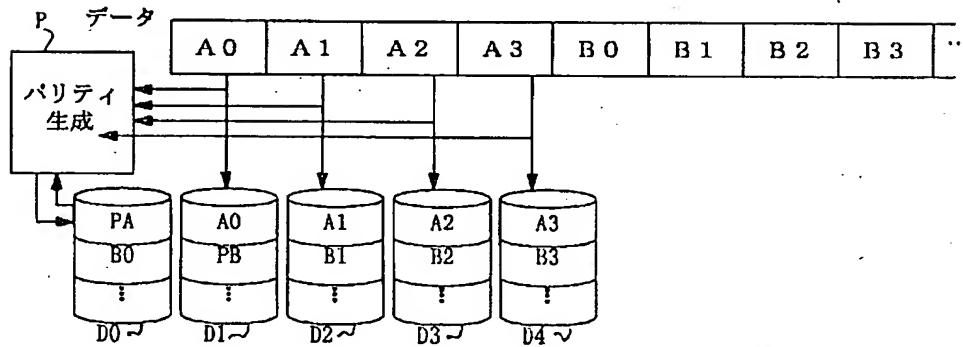
【図14】

RAID 4, 5を説明する図

(a) RAID4

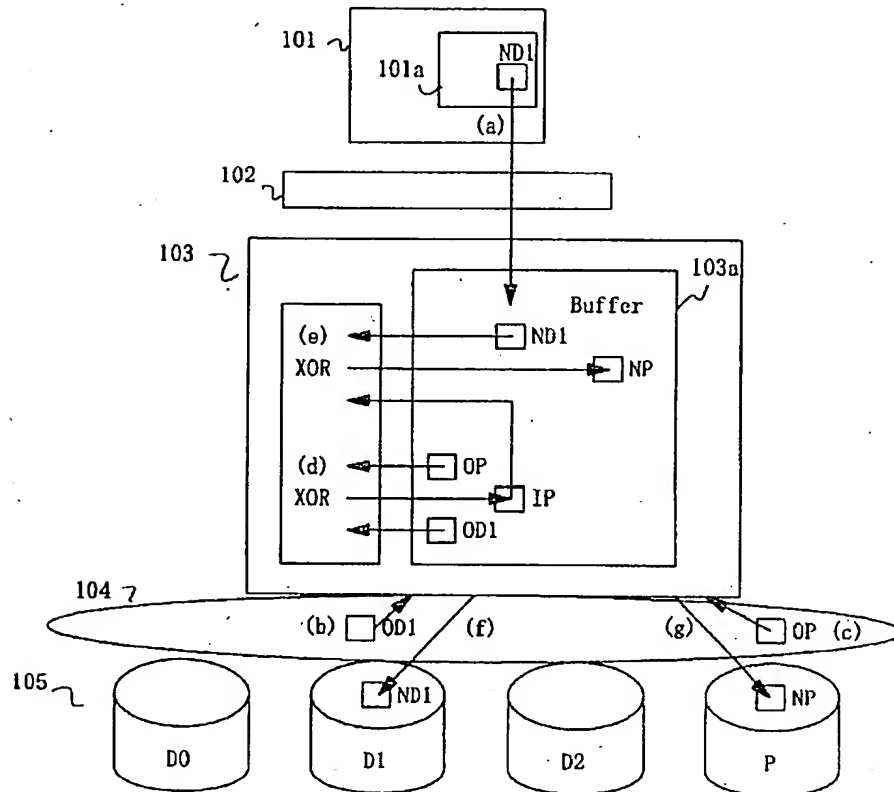


(b) RAID5



【図15】

ディスクアレイ装置における書き込み動作シーケンス
を示す図



フロントページの続き

(51) Int. Cl. ⁷	識別記号	F I	テーマコード(参考)
G 1 1 B 20/18	5 7 0	G, 1 1 B 20/18	5 7 0 Z
	5 7 2		5 7 2 B
			5 7 2 F
	5 7 6		5 7 6 Z

Fターム(参考) 5B018 GA02 KA15 KA21 LA06 MA14
QA05 QA06
5B065 BA01 CA11 CA30 CC08 CE12
EA02 EA23 EA24
5D044 BC01 CC04 DE68 DE94 IIII07
IIII17